Bayes Inference: Foundation of Machine Learning

Preface

To whom want to learn about Machine Learning, you may stuck at where to begin with, or you may get confused with the first chapter of a Machine Learning problem, like "Why the hell I am learning this?" or "Yes I understood the statistics you talked about here, but how is this related to Machine Learning, or How it may help my project?". So here we go, I hope we can discuss about the foundation of Machine Learning, Bayes Inference, and get a inuition of "How really Machine Learning works" at the end of this tutorial. Now, let's get started.

A glimpse on the core of Machine Learning

Before all, let's focus on "What Machine Learning really is" first. Machine Learning is developed to solve problems that can not be solved, or really poorly solved with ordinary algorithms. Machine Learning has nothing to do with the fancy "intelligence" in the movie, it can all go down to really simple problems.

For example, suppose we have a group of random variables: $(X_1, X_2, X_3, \ldots, X_N)$ observed, or unobserved. We want a model p_θ that capture all the relationship between variables. The approach of probabilistic generative models is to relate all variables by a learned **joint probability distribution** $p_\theta(X_1, X_2, \ldots, X_N)$. Intuitively, consider our p_θ, it should be a function that counts the impact of every point from $(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_N)$ on X_k. Is the function good enough? No. We neeeds to count the impact of correlation of (X_1, X_2) on X_k, also (X_1, X_2, X_3) on X_k. Adding all the correlation up, we will get exactly a **joint probablity distribution**, and that is what we are looking for.

Assume the variables were generated by some distribution $(X_1, \ldots, X_N)$ $p_\star(X)$. "Learning" the joint probability distribution, also called **density estimation** is the process choosing the parameters θ of a specified parametric joint distribution $p_\theta(X)$ to "best match" the "real" $p_\star(X)$.

To achieve the goal, we have three problems that we focusing to solve:

- How should we specify p_θ? or What should p_θ should look like? Is there any other way we can

represent the p_θ? (Because we would have infinite parameters brutal joint probablity distribution)

- How can we make sure that p_θ is the "Best Match"? or What is the meaning of "Best Match"?
- How can we find the best parameters of θ ?

Bayes Infernce Review

Now we see the role of conditional probablity plays in Machine Learning, and let's start with the Bayes Theorm.

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

A Probabilistic Perspective on Machine Learning Tasks

With representing the model by joint probability distribution, we can think about common machine learning tasks differently, where random variables represent:

- Input data X
- Discrete output or "Labels" C

  or

  Continous output Y

Then we have the joint probability distribution over these random variables,$p(C, X)$ or $p(X, Y)$, we see that this can be used for Machine Learning Tasks like:

1. Regression: $p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(X,Y)}{\int p(X,Y)dy}$
2. Classification / Clustering: $p(C|X) = \frac{p(X,C)}{\sum_C p(X,C)}$

Classification VS Clustering : Oberserved VS Unobserved Random Variables

The distinction between Classification VS Clustering, or Supervised vs Unsupervised Learning, is given by whether a random variable is **oberseved** or **unobserved**. For example:

Supervised Dataset:$\{x_i, c_i\}_{i=1}^N \sim p(X, C)$

In this case, the class labels are "observed", and we are looking for finding the conditional distribution $p(C|X)$ satisfies the supervised classification problem.

However, we may encounter datasets that only contains the "input data":

Supervised Dataset:$\{x_i\}_{i=1}^N \sim p(X, C)$

Notice that we did not change the generative assumption, our data x_i is still distributed according to a class label $C = c_i$, even though it is **unobserved** in the dataset. The common way to refer to an unobserved discrete class label is "cluster".

However, whether the label is observed or not, **it does not ultimately change our goal**, which is to have a model of the conditional distribution over the labels/clusters given the input data $p(C|X)$.